Growth Mixture Modeling with Measurement Selection

Abby Flynt and Nema Dean

Abstract

Growth mixture models are an important tool for detecting group structure in repeated measures data. Unlike traditional clustering methods, they explicitly model the repeat measurements on observations, and the statistical framework they are based on allows for model selection methods to be used to select the number of clusters. However, the basic growth mixture model makes the assumption that all of the measurements in the data have grouping information/separate the clusters. In other clustering contexts, it has been shown that including non-clustering variables in clustering procedures can lead to poor estimation of the group structure both in terms of the number of clusters and cluster membership/parameters. In this paper, we present an extension of the growth mixture model that allows for incorporation of stepwise variable selection based on the work done by Maugis et al. (2009) and Raftery and Dean (2006). Results presented on a simulation study suggest that the method performs well in correctly selecting the correct variables and improves on recovering the cluster structure compared with the basic growth mixture model. The paper also presents an application of the model to a clinical study dataset and concludes with a discussion and suggestions for directions of future work in this area.

Keywords: Cluster analysis, growth mixture model, repeated measurements, longitudinal data, measurement selection

1 Introduction

Cluster analysis is the search for group structure in multivariate data where little or no *a priori* information about groups is available (Everitt et al., 2011). There are a wide variety of different types of cluster analysis approaches available. These can be broadly categorized into three classes: algorithmic (which includes k-means (MacQueen, 1967) and hierarchical clustering (Ward, 1963)), non-parametric (mode-hunting, cluster tree approaches, (Wishart, 1969), Section 11 in (Hartigan, 1975) and (Hartigan, 1981)) and parametric (finite mixture model clustering (Titterington et al., 1985)). Finite mixture model clustering (also commonly known as model-based clustering) is becoming more popular in many application areas due to the wide availability of software, ease of interpretation of output and limited number of subjective decisions necessary for its application.

This paper will focus on growth mixture modeling which is a special case of finite mixture model clustering. Growth mixture modeling is a framework that allows for cluster detection in situations with repeated measurements. It was first introduced by Muthén and Shedden (1999) and a good review can be found in Ram and Grimm (2009). The growth mixture model (GMM) framework allows for modeling of the repeated measurements either directly or as a regression model of outcome measurements on explanatory measurements.

One assumption made by the GMM is that all of the repeated measurements are important to the group structure. Figure 1 (a) shows an example where this is the case. However, if some measurements do not vary across groups, we may have a situation similar to Figure 1 (b). In the latter example, only the last two of the repeated measurements are important for separating the three groups. So the assumption of all measurements having clustering information may not be true and if noise variables are included, it has been shown in other contexts (Raftery and Dean, 2006; Rusakov and Geiger, 2005) that this can be detrimental to the performance of the clustering method. It is also the case that from a substantive point of view, knowing which measurements/time points differentiate between groups may be of interest in itself. For example, Figure 1 (b) could be a case where a medication does have differing impacts on groups of patients but this difference

effect is delayed and not visible in the first two measurements/time points. This paper applies the variable selection method proposed in Raftery and Dean (2006) and Maugis et al. (2009) to the GMM to simultaneously allow for selection of the repeated measurements that drive the clustering with application of the clustering itself.

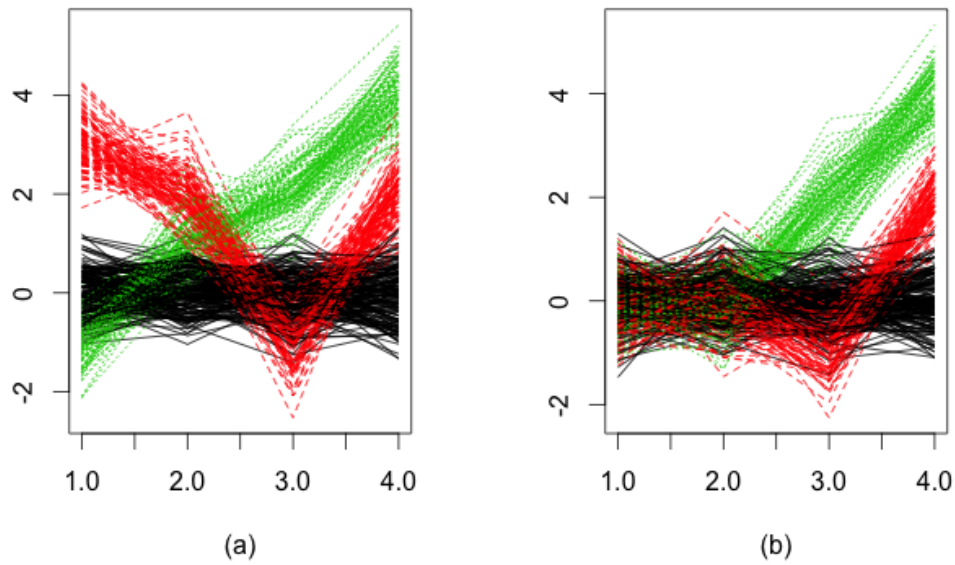

Figure 1: (a) 3 groups, where all time-points along x-axis are important for separating groups; (b) 3 groups, where only 3rd and 4th time-points are important for separating groups. Different groups have different line types and colours.

We begin with the methods in 2, where in Section 2.2, we introduce the general GMM and discuss its properties and estimation. We then summarize the variable selection framework of Raftery and Dean (2006) in Section 2.3. The specific variable selection for both a basic GMM (without cluster-specific measurement regression) and the regression GMM is explained in 2.4. These are applied in Section 3 to a variety of different settings in a simulation study, with the results presented in Section 3.1, followed by results on the Pittsburgh 600 dataset in Section 3.2. The paper wraps up with a summary of the main results, some caveats and future directions for further research on this topic in Section 4.

2 Methods

2.1 Finite Mixture Model Clustering

One of the first major analyses using a finite mixture model was introduced by Pearson (1894). Finite mixture model clustering assumes that instead of a single homogenous population, data may come from a heterogenous population made up of homogenous sub-populations. Rather than directly model the overall population with a single density, each sub-population is modeled with its own density. The overall population density can then be expressed as a weighted sum of the component/sub-population densities, with the weights set to the proportions each sub-population makes up of the overall population.

So, if we have an individual (multivariate) observation $\boldsymbol{y}$, the mixture density with K components/sub-populations is given by:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}|\boldsymbol{\theta}_k), \tag{1}$$

where

$$\pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1.$$

Here, the π_k's are the mixture proportions and the $\boldsymbol{\theta}_k$ are the sets of parameters for each component density. If $\boldsymbol{y}$ is continuous, then f_k is often taken to be Gaussian and then $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, a set of component specific mean vectors $\boldsymbol{\mu}_k$ and component specific covariance matrices $\boldsymbol{\Sigma}_k$. Clustering using finite mixture models with Gaussian component densities is commonly known as model-based clustering (see Fraley and Raftery (1998) for details).

If we have a single variable y that is related to another variable x or a vector of variables $\boldsymbol{x} = (x_1, \ldots, x_p)$, then we can cluster the relationship between y and $\boldsymbol{x}$ via a finite mixture of regression models as given by:

$$f(y|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y|\boldsymbol{x}, \boldsymbol{\theta}_k) \tag{2}$$

For continuous y, f_k is usually assumed to be Gaussian distributed, with $\theta_k = (\boldsymbol{\beta}_k, \sigma_k^2)$, a set of component specific regression parameter vectors $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ and component specific variance σ_k^2. The π_k's remains as before. Regression relationships which are more complex than linear can be used and for a non-continuous y, more complex models like generalized linear models can be used.

Estimation of the finite mixture model can be performed via either frequentist or Bayesian inferential methods. In the frequentist setting, the EM algorithm (Dempster et al., 1977) or variants thereof (Gupta and Chen, 2011; McLachlan and Krishnan, 2008) is commonly used for estimation, where the observed data is augmented by missing data, in this case an identifier of which mixture component each observation is generated from. Fraley and Raftery (2002) gives details of model-based clustering estimation via EM.

2.1.1 Choice of number of components

Choosing the number of mixture components that best fits the data can be posed as a model choice question for finite mixture model clustering methods, in contrast to traditional algorithmic methods. Each number of components defines a different model so we can score the fit of each model, then choose the model (and associated number of components) that scores best. One of the most commonly used scoring mechanisms is the Bayesian Information Criterion, BIC (Schwarz, 1978). For a particular model M with number of independent parameters ν and number of data points, n, the BIC can be defined as

$$BIC(M) = -\log(\text{maximized likelihood of model } M) + \nu \log(n) \tag{3}$$

This essentially looks at how well the model fits the data, via the maximized log likelihood and penalizes the model complexity by the number of parameters required, weighted by the log of the number of observations. The best model will be the one with the *smallest* BIC score.

2.1.2 Assignment of observations to components

For each observation y, instead of a hard assignment to a particular cluster, mixture model clustering can produce a vector of posterior class membership probabilities via Bayes' rule, using the fitted model parameters:

$$\hat{p}_{sm} = P(\text{component } = m | y_s, \boldsymbol{x}_s, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_m f_m(y_s | \boldsymbol{x}_s, \hat{\boldsymbol{\theta}}_m)}{\sum_{k=1}^{K} \hat{\pi}_k f_k(y_s | \boldsymbol{x}_s, \hat{\boldsymbol{\theta}}_k)} \tag{4}$$

These class membership probabilities are one of the advantages of mixture model clustering. They can give a measure of uncertainty in the assignment of observations to components, which is not available from a hard clustering. They can also give an indication of the degree of overlap between components.

If a hard classification is required, the *maximum a posteriori* (MAP) mapping can be used. This assigns an object to the mixture component that has highest value in the class membership probabilities. The MAP classification is simply:

$$\text{Component for subject } s = \arg\max_m \hat{p}_{sm}$$

2.1.3 Connecting a fitted mixture model to clustering

Once the best mixture model for the data has been selected, the most common assumption is that each component represents a cluster. Recent work by Baudry et al. (2010), Hennig (2010), Scrucca (2016), and Melnykov (2016) have examined different ways of combining components to create multi-component clusters. This paper assumes each component found represents a cluster but the GMM model could also use the ideas from these papers to merge components and create new clusters.

2.2 Growth Mixture Model

The growth mixture model is a mixture model applicable to data with multiple measurements, e.g. individuals recorded at multiple points over time. The assumption of conditional independence is made between different (sets of) repeated measurements, conditioned on component membership. This reduces the multivariate component density f_k to a product of univariate component densities. If we have S subjects/individuals, each with N_s repeated measurements, the k^{th} component density decomposes to the following product for multivariate observation $\boldsymbol{y}_s$:

$$f_k(\boldsymbol{y}_s|\boldsymbol{\theta}) = \prod_{n=1}^{N_s} f_{nk}(y_{sn}|\boldsymbol{\theta}_{nk}), \tag{5}$$

or for the case of repeated measurements on an outcome variable y and set of covariates $\boldsymbol{x}$ we have:

$$f_k(\boldsymbol{y}_s|\boldsymbol{x}_s, \boldsymbol{\theta}) = \prod_{n=1}^{N_s} f_{nk}(y_{sn}|\boldsymbol{x}_{sn}, \boldsymbol{\theta}_{nk}), \tag{6}$$

where n indexes the repeated measurements in each subject. This conditional independence model is similar to that used in latent class analysis. Note that equation (5), without covariates in the Gaussian setting, is a special case of the model-based clustering model with covariance parameterisation set to be a diagonal matrix within each component (where the diagonal elements are not required to be equal within or across components). This is the "VVI" in the `mclust` (model-based clustering package (C. Fraley and Scrucca, 2012; Fraley and Raftery, 2002) in the `R` software language (R Core Team, 2015)) parlance.

We can use the EM algorithm estimation along with the standard methods as discussed in Sections 2.1.1, 2.1.2 and 2.1.3 to produce a clustering model in this framework.

2.2.1 Mixed mode and missing data

The conditional assumption means that the outcome variables can be mixed mode data, i.e. of different types (continuous or categorical, binary versus count, etc.), as conditional independence

means that maximization within the M step of the EM algorithm takes place over each repeated measurement separately. The E step of the EM, essentially given by equation (4), can easily be adapted to this situation as well, since it involves evaluation of each density separately before the product is taken.

Similarly, if there is data missing for some of the repeated measurements for some subjects, the data from these subjects can still be used in the estimation of model parameters and component memberships can still be calculated on these subjects. The maximization for a particular repeated measurement can take place over only the subjects for which there is no missing data on that measurement as a result of the conditional independence assumption. Similarly, equation (4) can be updated in the following way:

$$\hat{p}_{sm} = P(\text{component } = m | y_s, \boldsymbol{x}_s, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_m f_m(y_s|\boldsymbol{x}_s, \hat{\boldsymbol{\theta}}_m)}{\sum_{k=1}^{K} \hat{\pi}_k f_k(y_s|\boldsymbol{x}_s, \hat{\boldsymbol{\theta}}_k)},$$

where

$$f_k(y_s|\boldsymbol{x}_s, \hat{\boldsymbol{\theta}}_k) = \prod_{n:\, y_{sn}, \boldsymbol{x}_{sn} \text{not missing}} f_{nk}(y_{sn}|\boldsymbol{x}_{sn}, \hat{\boldsymbol{\theta}}_{nk}).$$

2.3 Variable Selection Framework

The variable selection procedure proposed in Raftery and Dean (2006) and Dean and Raftery (2010) is ideal for application to the GMM. It reduces the variable selection problem to considering whether a single variable is useful for clustering or not. This is then combined with a search procedure (e.g. stepwise, backward, forward, headlong, etc.) to select a subset of the original variables as useful for clustering.

2.3.1 Variable Selection Comparison Models

To check if a proposal variable is useful for clustering, two opposing models are posited (useful for clustering versus not) and the fit compared to see which model has stronger evidence. For each stage we have three (potential) sets in a partition of the variables/measurements $\boldsymbol{y}$:

- $y^{(proposal)}$ - the (single) variable being proposed as a clustering variable
- $\boldsymbol{y}^{(current)}$ - the current set of selected clustering variables
- $\boldsymbol{y}^{(not\ selected)}$ - all other variables (not being proposed or currently selected as clustering variable)

. The model for $y^{(proposal)}$ being useful for clustering, M_1, is a product of two sub-models:

$$M_1(\boldsymbol{y}) = M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)}) \times M_{not\ clust}(\boldsymbol{y}^{(not\ selected)}|y^{(proposal)}, \boldsymbol{y}^{(current)}),$$

where M_{clust} indicates a clustering model was fitted to the set of variables in parentheses and $M_{not\ clust}$ is a non-clustering model.

The model for $y^{(proposal)}$ *not* being useful for clustering, M_2, is a product of three sub-models:

$$M_2(\boldsymbol{y}) = M_{clust}(\boldsymbol{y}^{(current)}) \times M_{not\ clust}(y^{(proposal)}|\boldsymbol{y}^{(current)}) \times M_{not\ clust}(\boldsymbol{y}^{(not\ selected)}|y^{(proposal)}, \boldsymbol{y}^{(current)}).$$

Different approaches were taken with respect to the sub-model, $M_{not\ clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$, for the relationship between the proposal variable and the current clustering variables in the model where the proposal variable does not have a clustering role.

1. In the original paper, Raftery and Dean (2006), where model-based clustering with Gaussian components was considered, this took the form of a linear regression of $y^{(proposal)}$ on the full set of current clustering variables $\boldsymbol{y}^{(current)}$.
2. In the Dean and Raftery (2010) paper, because conditional independence was assumed in the LCA model (Lazarsfeld and Henry, 1968) applied to the categorical variables under consideration, it made sense to assume complete independence of $y^{(proposal)}$ from $\boldsymbol{y}^{(current)}$. This results in a reduction of $M_{not\ clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$ to $M_{not\ clust}(y^{(proposal)})$.
3. A compromise between these two extremes was proposed by Maugis et al. (2009) where variable selection was run on the regression of $y^{(proposal)}$ on the full set of current cluster-

ing variables $\boldsymbol{y}^{(current)}$ (allowing $y^{(proposal)}$ to depend on all, a subset of or none of the set $\boldsymbol{y}^{(current)}$).

The modeling choices for the GMM case for $M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$, based on 3, are discussed further in Section 2.4.

For a proposed variable $y^{(proposal)}$, once models M_1 and M_2 have been estimated from the data, a method for evaluating the evidence for one versus the other is needed. The obvious choice for doing so is to examine the Bayes factor of model 1 versus model 2, B_{12}

$$B_{12} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)},$$

where $p(\boldsymbol{y}|M_i)$ is the integrated likelihood of model M_i. These integrated likelihoods are not available in closed form for finite mixture models. So instead, an approximation of the Bayes factor using the BIC is implemented.

$$\log(B_{12}) \approx BIC(M_1) - BIC(M_2). \tag{7}$$

As lower values of the BIC, as defined in equation (3), indicate a better fit, if the difference in BIC values in equation (7) is negative, this indicates more evidence in support of model M_1 than M_2, suggesting the proposal variable, $y^{(proposal)}$ is useful for clustering. Conversely, if the difference is positive, there is more evidence supporting M_2 than M_1, suggesting the proposal variable, $y^{(proposal)}$ is not useful for clustering. The natural default value of the threshold for the BIC difference for deciding if $y^{(proposal)}$ should be included in the set of clustering variables is 0. This value can be altered if stronger evidence is believed to be necessary before inclusion of a proposal variable into the set of clustering variables.

For M_1 we have:

$$BIC(M_1) = BIC(M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)}) + BIC(M_{not\,clust}(\boldsymbol{y}^{(not\,selected)}|y^{(proposal)}, \boldsymbol{y}^{(current)}))$$

For M_2 we have:

$$\begin{aligned} BIC(M_2) &= BIC(M_{clust}(\boldsymbol{y}^{(current)})) + BIC(M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)}) \\ &\quad + BIC(M_{not\,clust}(\boldsymbol{y}^{(not\,selected)}|y^{(proposal)}, \boldsymbol{y}^{(current)})) \end{aligned}$$

We see that $M_{not\,clust}(\boldsymbol{y}^{not\,selected}|y^{(proposal)}, \boldsymbol{y}^{(current)})$ appears in both models so this cancels in the difference giving us:

$$\begin{aligned} BIC_{diff}(y^{(proposal)}) &= BIC(M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)})) \\ &\quad - \Big(BIC(M_{clust}(\boldsymbol{y}^{(current)})) + BIC(M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})\Big) \end{aligned} \tag{8}$$

So, for each proposed, $y^{(proposal)}$, the cluster model is fit on the set of variables including the current set of clustering variables $\boldsymbol{y}^{(current)}$ along with the proposal variable $y^{(proposal)}$ and the BIC score for that model, $BIC(M_1)$, is calculated. The cluster model is fit on only the current set of clustering variables $\boldsymbol{y}^{(current)}$ and the BIC score for this model is calculated, $BIC(M_{clust}(\boldsymbol{y}^{(current)}))$, and either the regression (with or without variable selection) is fit for $y^{(proposal)}$ on $\boldsymbol{y}^{(current)}$ or a single component cluster model is fit on $y^{(proposal)}$ and $BIC(M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$ or $BIC(M_{not\,clust}(\boldsymbol{y}^{(current)}))$ is produced. These are plugged into the $BIC_{diff}(y^{(proposal)})$ to give a number that can be used to make a decision. This is then combined with a search algorithm, such as those presented in the next section, to produce a set of selected clustering variables.

2.3.2 Variable Selection Search Algorithm

Given the nature of the data examined where variables correspond to repeated measurements or time points, the number of variables is not expected to be large. This means that there will be no issues with fitting a GMM using the full set of measurements. Therefore we describe two backward search algorithms that will be used in producing the results for Section 3.

Basic Greedy Backward Search

The standard greedy backward search proceeds as follows:

1. Start with all variables in the $\boldsymbol{y}^{(current)}$ set
2. Take each variable from $\boldsymbol{y}^{(current)}$ individually in turn as $y^{(proposal)}$:
    - Fit models M_1 and M_2
    - Calculate BIC_{diff} using equation (8)
3. Choose the variable with largest BIC_{diff} value
4. If the variable's BIC_{diff} is larger than the chosen threshold (usually 0), then remove this variable from the set of clustering variables and return to step 2. Otherwise, halt the algorithm.

Greedy Backward Search with Monotonicity

Given that one of the common forms of data to which GMM is applied is repeated measurements where there is a temporal ordering to the recording of the repeated measurements, a greedy backward search with monotonicity may be of interest.

This type of search proceeds as follows:

1. Start with all variables in the $\boldsymbol{y}^{(current)}$ set
2. Take the *earliest* and the *latest* variable from $\boldsymbol{y}^{(current)}$ individually in turn as $y^{(proposal)}$:
    - Fit models M_1 and M_2
    - Calculate BIC_{diff} using equation (8)
3. Choose the variable with largest BIC_{diff} value
4. If the variable's BIC_{diff} is larger than the chosen threshold (usually 0), then remove this variable from the set of clustering variables and return to step 2. Otherwise, halt the algorithm.

The backward greedy search in the case of limited number of variables/measurements should be the most efficient (where there is at least some clustering going on). However, it is perfectly

feasible to look at forward or forward-and-backward stepwise search types as well as headlong (as described in Dean and Raftery (2010)) instead of greedy approaches as wrappers for the framework from Section 2.3.1.

2.4 Growth Mixture Model with Variable Selection

This section presents the details for model $M_1(\mathbf{y})$, where $y^{(proposal)}$ is useful for clustering and $M_2(\mathbf{y})$, where $y^{(proposal)}$ is *not* useful for clustering in the framework of growth mixture modeling with variable selection. As discussed in Section 2.3.1, we can ignore $M_{not\,clust}(\boldsymbol{y}^{not\,selected}|y^{(proposal)}, \boldsymbol{y}^{(current)})$ as it appears in both models. We focus on $M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)})$, $M_{clust}(\boldsymbol{y}^{(current)})$ and $M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$. Both M_{clust} models come directly from Equations (5) and (6) depending on whether or not we have the inclusion of covariates.

Let $\mathcal{V}$ be the set of indices from N_s for the current clustering variables. For example, if there were 5 repeated measurements and the third and fifth are current clustering variables, then $\mathcal{V} = \{3, 5\}$. Additionally, let p represent the index of the proposed clustering variable.

Without covariates, we have

$$M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k \prod_{v \in \mathcal{V} \cup p} f_{vk}(y_{sv}|\boldsymbol{\theta}_{vk}),$$

and

$$M_{clust}(\boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k \prod_{v \in \mathcal{V}} f_{vk}(y_{sv}|\boldsymbol{\theta}_{vk}).$$

For both of these clustering models, f_{vk} is Gaussian with $\boldsymbol{\theta}_{vk} = (\boldsymbol{\mu}_{vk}, \boldsymbol{\Sigma}_{vk})$.

The addition of covariates gives us

$$M_{clust}(y^{(proposal)}, \boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k \prod_{v \in \mathcal{V} \cup p} f_{vk}(y_{sv}|\boldsymbol{x}_{sv}, \boldsymbol{\theta}_{vk}),$$

and

$$M_{clust}(\boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} \sum_{k=1}^{K} \pi_k \prod_{v \in \mathcal{V}} f_{vk}(y_{sv}|\boldsymbol{x}_{sv}, \boldsymbol{\theta}_{vk}),$$

where again, f_{ck} is Gaussian, but for this finite mixture of regression models, $\boldsymbol{\theta}_{ck}$ is a set of component specific regression parameters as described in Section 2.1.

The model for $M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)})$ without covariates is given by

$$M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} f_p(y_s^{(proposal)}|\boldsymbol{y}_s^{(current)*}, \boldsymbol{\theta}_p),$$

where $\boldsymbol{y}^{(current)*}$ is a selection of $\boldsymbol{y}^{(current)}$ chosen using a regression variable selection based on BIC (Raftery, 1995).

The addition of covariates requires us to first fit the regression of $y^{(proposal)}$ on $\boldsymbol{x}^{(proposal)}$, we then select $\boldsymbol{y}^{(current)*}$ from the residuals of the previous regression on $\boldsymbol{y}^{(current)}$. Giving a final model of

$$M_{not\,clust}(y^{(proposal)}|\boldsymbol{y}^{(current)}) = \prod_{s=1}^{S} f_p(y_s^{(proposal)}|\boldsymbol{x}_s^{(proposal)}, \boldsymbol{y}_s^{(current)*}, \boldsymbol{\theta}_p).$$

3 Results

The starting values for EM estimation in all the following sections were produced using the k-means algorithm on all $\boldsymbol{y}$ values for the measurements being considered in the model. This was found to produce reasonable sets of starting values in general and was computationally quick. Other starting value schemes could, of course, be incorporated.

Only complete data (no missing values) were generated in the simulations in Section 3.1 and complete cases were used in the dataset in Section 3.2. The presented methodology can easily be extended to include missing values using the ideas in Section 2.2.1, but evaluating the impact of this was not considered to be the goal of this paper.

3.1 Simulation Study

For each simulated data set, we compare the final GMM model after variable selection with the GMM model using all of the variables. We can compare the accuracy of clustering by looking at the difference in the Adjusted Rand Index (ARI, Hubert and Arabie (1985)) compared to the true simulated partition for each model. Additionally, we can compare the accuracy of estimation by looking at the difference in the Root Mean Square Error (RMSE, Steel and Torrie (1960)) for each model. We include the RMSE for the full set of variables based on the model using the selected variables and the model using all variables. Note that the RMSE for the model estimated using selected variables includes estimation of the non-selected clustering variables assuming a 1 group model. We also provide results on the number of non-clustering variables chosen and the number of clusters chosen by each model.

All results are for 20 time points with 3 groups. The non-clustering time points all have intercept 0 and slope 1 with standard deviation 0.5 and standard normal distributed explanatory variables. All simulations are repeated 50 times and average results are reported in Tables 1, 2, 3 and 4.

Tables 1, 2 and 3 all refer to clusterings with roughly equal sized clusters (in terms of membership/mixture probabilities) with varying degrees of separation in the two clustering time points. In the first simulation summarized in Table 1, the first clustering variable has slope parameters (1, 3, -2) and the second clustering variable has slope parameters (1, 2.5, -0.5) over the three groups respectively. 98% of the simulations correctly selected both the clustering variables and 43% of the simulations correctly chose 3 groups. Additionally, 78% of the simulations had higher ARI and 66% of the simulations had lower RMSE for the model using only the selected clustering variables. In the second and third simulation, both clustering variables share the same slopes for the three groups, with Table 2 having slopes (1, 2.5, -0.5) and Table 3 having slopes (1, 3, -2). We seen an improvement in all of our measures of evaluation for these two simulations.

Finally in Table 4, the membership/mixture probabilities place 70% of the observations in one group and split the rest into the other two groups. The slopes for the clustering variables match

that of the third simulation. Even with unbalanced groups, the clustering variables are selected correctly 96% of the time and three groups are chosen 56% of the time. It is worth noting that clustering of all the time points resulted in only 2 groups being chosen for every repetition of the simulation. The ARI was higher for 72% of the simulations and the RMSE was smaller for 90% of the simulations for the model using only the selected clustering variables.

Overall, as can be seen from the tables, the variable selection procedure almost always selects the clustering variables, but also often includes some additional variables. However, as can be seen from the ARI, in terms of group structure recovery, most of the time it is better to just select the cluster model on the basis of the selected variables rather than the full variable set. Similarly, if good estimation performance is the goal, again in terms of RMSE, it is usually better to fit the model based on the selected variables alone.

3.2 Application

In this section, we apply variable selection to the Pittsburgh 600, a longitudinal data set composed of five depression studies, each with a different treatment protocol (Thase et al. (1997)). Subjects were diagnosed as clinically depressed upon entrance to the study and there are 26 weekly scores indicating their current level of depression. In addition to their score, we have several explanatory variables, including the subjects' gender, age and medication status.

Because most trajectories have intermittent missingness, we have taken only the complete trajectories for a subset of four time points. There were 74 subjects that had complete observations at Weeks 1, 9, 14 and 25 of the study. Figure 2 shows the trajectories of these subjects over time. We can see that overall there is a decreasing trend, showing a reduction in clinical depression scores.

Using subject age as the explanatory variable, variable selection chose Week 14 as the only time point that is useful for clustering. Clustering on Week 14 produces two classes, containing 46 and 28 trajectories as seen in Figure 3. In Figure 4, we see that the two classes obtained from growth mixture model are really driven by different intercepts.

Simulation parameters									
Group mixture weights	Group 1 = 0.3			Group 2 = 0.3			Group 3 = 0.4		
Clustering time points	5 and 15								
Clustering time points slopes	Group 1			Group 2			Group 3		
Time point 5	1			3			-2		
Time point 15	1			2.5			-0.5		
Simulation results									
Clustering time point selected	Time point 5			Time point 15			Both time points		
% simulations	98%			100%			98%		
Number of non-clustering time points selected	0	1	2	3	4	6	7	8	17
% simulations	26%	26%	6%	8%	8%	8%	12%	4%	2%
Number of groups chosen	2	3	4						
% of simulations for model with *selected* variables	14%	42%	44%						
% of simulations for model with *all* variables	70%	12%	18%						

Summary statistics for difference between ARI for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.4969	0.01145	0.15330	0.18270	0.30260	0.7489
78% of simulations had a higher ARI for the model using only the selected variables					

Summary statistics for difference between RMSE for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.64120	-0.22790	-0.11790	-0.12330	0.04693	0.92630
66% of simulations had lower RMSE for all time points for the model using only the selected variables					

Table 1: First simulations set

Simulation parameters									
Group mixture weights	Group 1 = 0.3			Group 2 = 0.3			Group 3 = 0.4		
Clustering time points	5 and 15								
Clustering time points slopes	Group 1			Group 2			Group 3		
Time point 5	1			2.5			-0.5		
Time point 15	1			2.5			-0.5		
Simulation results									
Clustering time point selected	Time point 5			Time point 15			Both time points		
% simulations	100%			100%			100%		
Number of non-clustering time points selected	0	1	2	3	4	5	6	14	17
% simulations	66%	20%	6%	2%	1%	1%	2%	1%	1%
Number of groups chosen	2	3	4						
% of simulations for model with *selected* variables	19%	52%	29%						
% of simulations for model with *all* variables	94%	6%	0%						

Summary statistics for difference between ARI for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.428100	-0.002086	0.063560	0.050580	0.130200	0.275100
77% of simulations had a higher ARI for the model using only the selected variables					

Summary statistics for difference between RMSE for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.5284	-0.3689	-0.3232	-0.2634	-0.1757	0.3829
94% of simulations had lower RMSE for all time points for the model using only the selected variables					

Table 2: Second simulations set

Simulation parameters								
Group mixture weights	Group 1 = 0.3		Group 2 = 0.3			Group 3 = 0.4		
Clustering time points	5 and 15							
Clustering time points slopes	Group 1		Group 2			Group 3		
Time point 5	1		3			-2		
Time point 15	1		3			-2		
Simulation results								
Clustering time point selected	Time point 5		Time point 15			Both time points		
% simulations	100%		100%			100%		
Number of non-clustering time points selected	0	1	2	3	4	5	6	7
% simulations	45%	23%	7%	7%	7%	2%	5%	5%
Number of groups chosen	2	3	4					
% of simulations for model with *selected* variables	2%	77%	20%					
% of simulations for model with *all* variables	61%	34%	5%					

Summary statistics for difference between ARI for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.15790	-0.01606	0.17880	0.13660	0.21700	0.55370
73% of simulations had a higher ARI for the model using only the selected variables					

Summary statistics for difference between RMSE for clustering with selected variables versus clustering with all variables					
Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-1.143000	-0.531900	-0.479300	-0.388600	-0.047100	-0.005398
100% of simulations had lower RMSE for all time points for the model using only the selected variables					

Table 3: Third simulations set

Simulation parameters							
Group mixture weights	Group 1 = 0.7		Group 2 = 0.15			Group 3 = 0.15	
Clustering time points	5 and 15						
Clustering time points slopes	Group 1		Group 2			Group 3	
Time point 5	1		2.5			-0.5	
Time point 15	1		2.5			-0.5	
Simulation results							
Clustering time point selected	Time point 5		Time point 15			Both time points	
% simulations	96%		96%			96%	
Number of non-clustering time points selected	0	1	2	3	4	6	17
% simulations	54%	22%	10%	6%	4%	2%	2%
Number of groups chosen	1	2	3	4			
% of simulations for model with *selected* variables	4%	20%	56%	20%			
% of simulations for model with *all* variables	0%	100%	0%	0%			

Summary statistics for difference between ARI for clustering with selected variables versus clustering with all variables

Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.199900	-0.001656	0.106100	0.117800	0.211500	0.609400

72% of simulations had a higher ARI for the model using only the selected variables

Summary statistics for difference between RMSE for clustering with selected variables versus clustering with all variables

Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-0.27330	-0.15030	-0.10200	-0.09932	-0.05620	0.08910

90% of simulations had lower RMSE for all time points for the model using only the selected variables

Table 4: Fourth simulations set

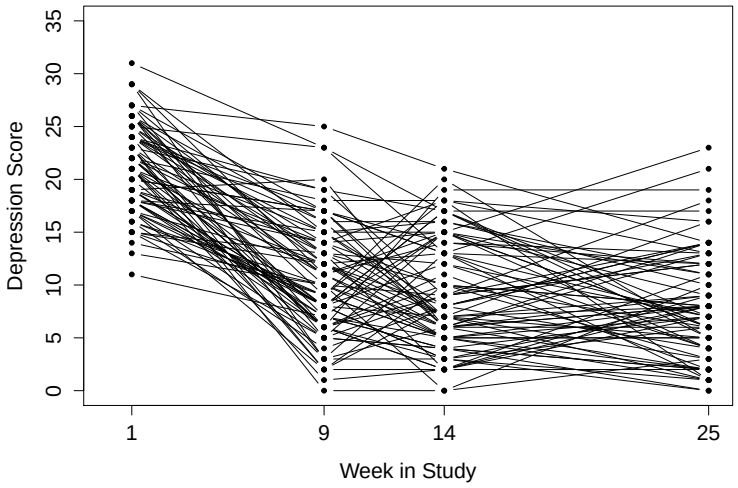

Figure 2: Subjects in Pittsburgh 600

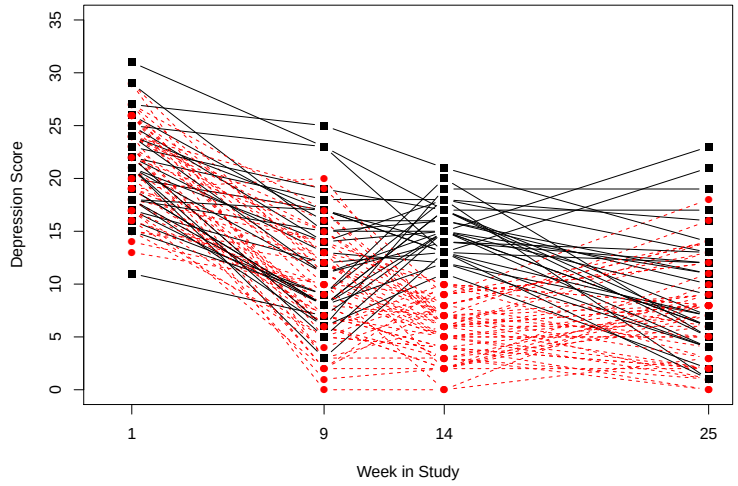

Figure 3: Clusters resulting from GMM fit only to Week 14

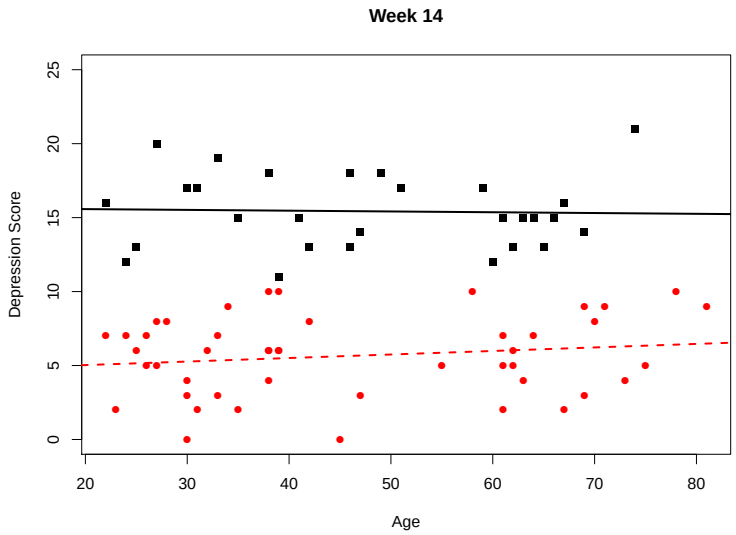

Figure 4: Clusters at Week 14

When a growth mixture model is fit to all 4 time points, again 2 groups of size 43 and 31 are selected. The clustering can be seen in Figure 5. When we compare these two clustering solutions, the ARI is 0.487 indicating that the solutions are not very similar. It is worth noting that we are seeing nearly identical results from the variable selection, growth mixture model procedure when the explanatory variable is changed to whether or not each subject is on a medication and also which of the 5 studies they belong to. Using the all of the time points for the growth mixture model with either of these explanatory variables however, produces more variation in the clustering.

4 Discussion

This paper presented a framework to incorporate variable selection into the growth mixture model for the settings where one is clustering repeated measurements over time in addition to possible covariates. The simulated results presented suggest from both an estimation and interpretation point of view it is preferable to apply variable selection when using a GMM. It results in better estimation of the number of groups present in the data and superior estimation of the parameters in the resulting regressions.

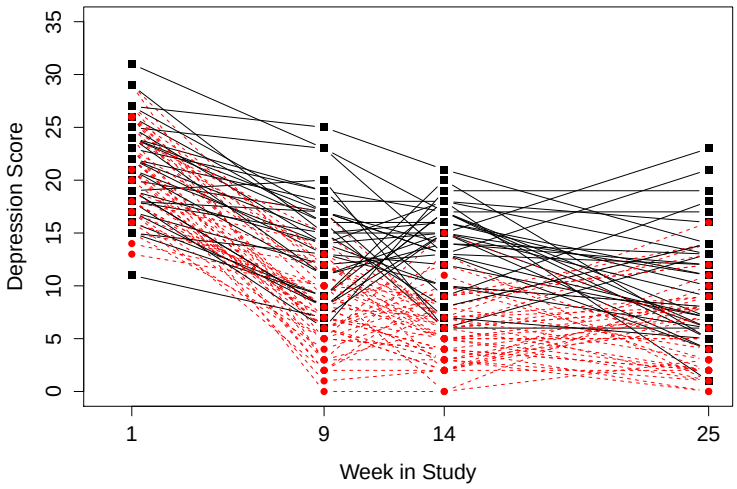

Figure 5: Clusters resulting from GMM fit to all four time points

The clustering found using the selected variables can often vary from that found using the full original set of variables/measurements as seen in the application in Section 3.2. Which cluster model is of more interest can depend on the context of the problem. If one is interested in the grouping on all variables, a priori, then the variable selection is interesting only in the sense of highlighting variables that best separate the groups, not in the sense of estimating the clustering itself. If this is so, then variable selection can be performed to find these variables but the clustering model estimation is best done on the full set. Or indeed an alternative post-hoc variable selection method could be used after the clustering has already been performed. In other situations, it may be the case that both variable selection and clustering on the selected variables is the goal. It may also still be of interest to compare the clustering found on the selected variables versus the clustering on the original full set of variables.

The framework presented here is extremely flexible and can easily be extended to facilitate estimation using incomplete observations as mentioned previously.

This framework could also be adapted to the generalized linear model framework in addition to the linear model approach presented in this paper, or the generalized additive model clustering case as seen in James and Sugar (2003). It would also be easy to extend the model to allow an

autoregressive time dependence instead of the conditional independence presented, either by using previous time points as covariates in the covariate GMM setting or using the framework presented in McNicholas and Subedi (2012).

In addition to variable selection in the clustering sense, it would also be possible to incorporate variable selection in the regression for the outcome variables on their covariates in the GMM setting.

A classification or supervised version of our methodology would be easy to implement since it essentially involves an observed version of the missing data (group labels) in the EM and a single M step for parameter estimation followed by a single E step for estimating group membership probabilities for new data. A hybrid semi-supervised approach, similar to Murphy et al. (2010), could also be considered.

Issues with this method can arise with difficulties in finding good starting values for the clustering. K-means does a good job in most cases but other methods could be used in place of this should problems arise. There can also be issues when a cluster is assigned to only one or a pair of trajectories as this can result in singular matrix issues during estimation of the cluster and regression parameters. Including a noise component in the clustering might be a possible method to deal with this issue (as a single observation cluster could be argued to be an outlier/noise observation). Bayesian estimation used as an alternative to the EM based estimation advocated in this paper could also help with regularization issues.

This paper presents an initial look at variable selection in the growth mixture model framework which will hopefully stimulate interesting further research in this area. R code for fitting the models described in this paper is available on request from the authors.

References

Baudry, J., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010), "Combining Mixture Components for Clustering." *Journal of Computational and Graphical Statistics*, 332–353.

C. Fraley, A. E. Raftery, T. B. M. and Scrucca, L. (2012), "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation," Tech. Rep. 597, Department of Statistics, University of Washington.

Dean, N. and Raftery, A. E. (2010), "Latent class analysis variable selection," *Annals of the Institute of Statistical Mathematics*, 62, 11–35.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011), *Cluster Analysis*, Wiley Series in Probability and Statistics, Chichester, UK: Wiley.

Fraley, C. and Raftery, A. E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The computer journal*, 41, 578–588.

— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 611–631.

Gupta, M. R. and Chen, Y. (2011), "Theory and Use of the EM Algorithm," *Foundations and Trends® in Signal Processing*, 4, 223–296.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley.

— (1981), "Consistency of single linkage for high-density clusters," *Journal of the American Statistical Association*, 76, 388–394.

Hennig, C. (2010), "Methods for merging Gaussian mixture components," *Advances in Data Analysis and Classification*, 3–34.

Hubert, L. and Arabie, P. (1985), "Comparing partitions," *Journal of classification*, 2, 193–218.

James, G. M. and Sugar, C. A. (2003), "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, 98, 565–576.

Lazarsfeld, P. F. and Henry, N. W. (1968), *Latent structure analysis*, Houghton Mifflin.

MacQueen, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009), "Variable Selection for Clustering with Gaussian Mixture Models," *Biometrics*, 65, 701–709.

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, Wiley.

McNicholas, P. and Subedi, S. (2012), "Clustering gene expression time course data using mixtures of multivariate t-distributions," *Journal of Statistical Planning and Inference*, 5.

Melnykov, V. (2016), "Merging mixture components for clustering through pairwise overlap," *Journal of Computational and Graphical Statistics*, 24, 66–90.

Murphy, T. B., Dean, N., and Raftery, A. E. (2010), "Variable Selection and Updating In Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications," *Annals of Applied Statistics*, 4, 396–421.

Muthén, B. and Shedden, K. (1999), "Finite mixture modeling with mixture outcomes using the EM algorithm," *Biometrics*, 55, 463–469.

Pearson, K. (1894), "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London, Series A*, 71.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. (1995), "Bayesian model selection in social research (with Discussion)," *Sociological Methodology*, 111–196.

Raftery, A. E. and Dean, N. (2006), "Variable selection for model-based clustering," *Journal of the American Statistical Association*, 101, 168–178.

Ram, N. and Grimm, K. J. (2009), "Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups," *International Journal of Behavioral Development*, 33, 565–576.

Rusakov, D. and Geiger, D. (2005), "Asymptotic Model Selection for Naive Bayesian Networks," *Journal of Machine Learning Research*, 6, 1–35.

Schwarz, G. E. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Scrucca, L. (2016), "Identifying connected components in Gaussian finite mixture models for clustering," *Computational Statistics & Data Analysis*, 93, 5–17.

Steel, R. G. D. and Torrie, J. (1960), *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*, McGraw Hill.

Thase, M. E., Greenhouse, J. B., Frank, E., Reynolds, C. F., Pilkonis, P. A., Hurley, K., Grochocinski, V., and Kupfer, D. J. (1997), "Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations," *Archives of General Psychiatry*, 54, 1009–1015.

Titterington, D. M., Smith, A. F., Makov, U. E., et al. (1985), *Statistical analysis of finite mixture distributions*, vol. 7, Wiley New York.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236–244.

Wishart, D. (1969), "Mode analysis: A generalization of nearest neighbor which reduces chaining effects," in *Numerical Taxonomy*, ed. Cole, A. J., Academic Press, pp. 282–311.